Knowledge-Based Systems xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Identifying the semantic orientation of terms using S-HAL for sentiment analysis

Q1 Tao Xu [a,b], Qinke Peng [a,b,*], Yinzhao Cheng [b]

[a] MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China
[b] Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 2 September 2011
Received in revised form 7 April 2012
Accepted 8 April 2012
Available online xxxx

Keywords:
Opinion mining
Q3 Sentiment analysis
Semantic orientation
Semantic polarity
Text mining

ABSTRACT

Sentiment analysis continues to be a most important research problem due to its abundant applications. Identifying the semantic orientation of subjective terms (words or phrases) is a fundamental task for sentiment analysis. In this paper, we propose a new method for identifying the semantic orientation of subjective terms to perform sentiment analysis. The method takes a classification approach that is based on a novel semantic orientation representation model called S-HAL (Sentiment Hyperspace Analogue to Language). S-HAL basically produces a set of weighted features based on surrounding words, and characterizes the semantic orientation information of words via a specific feature space. Because the method incorporates the idea underlying HAL and the hypothesis verified by the method of semantic orientation inference from pointwise mutual information (SO-PMI), it can quickly and accurately identify the semantic orientation of terms without the use of an Internet search engine. The results of an empirical evaluation show that our method outperforms other known methods.

1. Introduction

Q4 Sentiment analysis has received much attention over the past few years as a method of extracting useful knowledge from opinionated text (text with opinions or sentiments) [1–11]. It aims to automatically detect subjective information contained in text, identify the sentiment polarity of this subjective information, and estimate the strength of the sentiment polarity [5,7,12]. On the basis of sentiment analysis techniques, a variety of knowledge-based application systems have been developed recently [13–16]. An important research direction in sentiment analysis is to identify the sentiment polarity of individual words, known as words semantic orientation (referred to below as WSO), which indicates the evaluative character of a word and can vary in both direction (positive or negative) and intensity (mild to strong) [17–20]. A positive semantic orientation denotes a positive evaluation (i.e., praise) whereas a negative semantic orientation denotes a negative evaluation (i.e., criticism). Accurate identification of WSO is of great importance as it contributes to solve some key tasks in sentiment analysis, including detection of the subjectivity and semantic orientation of a whole text, and computation of the strength of the semantic orientation. Obviously, the technology for accurately identifying the WSO can benefit a variety of applications, ranging from automatic analysis of survey response to open questions, filtering "flame" for newsgroups, tracking voters' opinions about political candidates, developing personalized recommendation systems and intelligent automated chat systems [5,19].

Existing research into identifying the WSO falls into two rough categories: sentiment lexicon construction and automatic identification approach development. The representative work on sentiment lexicon construction includes the Lasswell Value Dictionary [21] and the General Inquirer Dictionary [22]. Recently, due to the foundational role of WordNet in semantic analysis [23], Esuli and Sebastiani constructed the SentiWordNet, which is specifically aimed at characterizing the sentiment polarity of word senses by learning methods [24,25]. These lexicons have a high accuracy, but due to their small coverage, they cannot be used to identify the semantic orientation of new cyber-words, phrases, and domain-dependent words. Automatic identification of WSO usually exploits statistical methods or learning methods to automatically produce the semantic orientation of a given word. The automatic identification approaches have a broad coverage, and so are able to satisfy a wide range of applications. Representative work on automatic identification of WSO mainly consists of: an algorithm proposed by Hatzivassiloglou and McKeown that can partition a set of adjectives into two clusters (positive and negative) by constructing a synonym and antonym connection graph [17]; a strategy presented by Turney and Littman for estimating the semantic orientation of a given word by computing the pointwise mutual information (PMI) between the given word and paradigm words [18,19]; an approach developed by Kamps et al. to identify the semantic orientation of a given word by detecting the synonym

Q2 * Corresponding author at: MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China. Tel./fax: +86 29 82667964.
E-mail addresses: txu@mail.xjtu.edu.cn (T. Xu), qkpeng@mail.xjtu.edu.cn (Q. Peng).

0950-7051/$ - see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.knosys.2012.04.011

relation based on WordNet [26]; a study by Esuli and Sebastiani suggesting that a classifier can be trained to identify the semantic orientation of a given word by utilizing an online glossary and standard machine learning algorithms [12]; and a proposal by Du and Tan to determine the semantic orientation of a given word by using spectrum optimization to detect communities on the semantic relation network [27].

Of all the above-mentioned methods for automatic identification of WSO, the most effective one is probably that of Semantic Orientation inference from pointwise mutual information (SO-PMI) proposed by Turney and Littman [18,19]. The SO-PMI method is regarded as the most effective approach in existing research on automatic semantic orientation identification because it possesses some prominent advantages, such as the highest accuracy in all available automatic identification methods, wide applicability to various types of identification objects (words or phrases), and easy implementation. However, a massive corpus, such as an Internet search engine, is crucial for guaranteeing the identification accuracy of the SO-PMI method. To obtain high accuracy, SO-PMI usually requires online support over the Internet, and therefore the identification speed is greatly restricted. In addition, even in the case where the largest currently available corpus, i.e., an Internet search engine, is adopted, the SO-PMI method still cannot achieve completely satisfactory identification accuracy.

The purpose of this study is to develop a rapid and precise method that can be used to automatically identify WSO without the online support of the Internet. To achieve this goal, we first present a novel semantic orientation representation model in which the semantic orientation information of words is characterized by a specific vector space. On the basis of the representation model, a classifier is then trained to identify the semantic orientation of terms (words or phrases). The presented semantic orientation representation model draws on work from two streams of research. One is the hypothesis that the semantic orientation of a word tends to correspond to that of its co-occurring neighbors, which is tested by the SO-PMI method. The second is the Hyperspace Analogue to Language (HAL) model that was proposed as an implementation of semantic space by Lund and Burgess [28,29]. In our work, a specific HAL model is constructed using a set of words with definite sentiment polarities as the base-space of HAL. According to the hypothesis tested by the SO-PMI method, we deem that this specific HAL model can be used as an implementation of a semantic orientation representation model. We refer to this specific HAL model as Sentiment HAL (S-HAL). By querying S-HAL, the semantic orientation feature vector of a word can be acquired, and the semantic orientation feature vector of a phrase can also be produced through a heuristic computation based on S-HAL. On the basis of the semantic orientation feature vector, a binary support vector machine (SVM) classifier is trained to predict the semantic orientation of any given terms.

In order to evaluate the proposed method, we conduct three different experiments. The first experiment is a comparison of the proposed method with other known methods. In the second experiment, we analyze the effect of model parameters' configuration on the identification performance. The final experiment mainly evaluates the heuristic combination method for producing semantic orientation feature vectors for phrases, which is crucial for identifying the semantic orientation of phrases.

The rest of the paper is organized as follows. Section 2 introduces some related work that constitutes the basis of our research. Section 3 presents the S-HAL model and a method for automatic semantic orientation identification based on the model. Section 4 contains the experimental evaluation of the proposed identification method, and Section 5 concludes this paper with some ideas for future work.

2. Related work

This section briefly introduces some related work on the SO-PMI algorithm and the HAL model, which constitutes an important basis for our work.

2.1. SO-PMI

Turney and Littman proposed a general strategy for identifying the semantic orientation of a term according to its statistical association with a set of positive and negative paradigm words [19]. This general strategy was called SO-A (Semantic Orientation from Association). The basic idea of the strategy can be summarized as follows.

Given two minimal sets of paradigm words with predefined positive and negative semantic orientations, the semantic orientation of a word w, denoted by SO-A(w), can be computed from the strength of its association with the paradigm words. This is written as:

$$\text{SO-A}(w) = \sum_{p_w \in S_p} A(w, p_w) - \sum_{n_w \in S_n} A(w, n_w)$$

where S_p denotes the set of paradigm words with a positive semantic orientation, S_n denotes the set of paradigm words with a negative semantic orientation, and A denotes any measurement for association between two words. Obviously, selecting a particular A can lead to particular instances of the strategy.

In Turney and Littman's work, S_p and S_n were defined, respectively, as follows:

S_p = {good, nice, excellent, positive, fortunate, correct, superior},
S_n = {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

In order to construct a reasonable A, Turney and Littman examined several approaches and found that the most effective was a method based on computing PMI. The PMI between two words, w and w_i, is calculated by:

$$\text{PMI}(w, w_i) = \log_2\left(\frac{\Pr(w, w_i)}{\Pr(w)\Pr(w_i)}\right)$$

Turney and Littman also proposed to calculate approximate values of PMI using the NEAR operator of the AltaVista search engine. The NEAR operator produces a match for a document when its operands appear in the document at a maximum distance of ten words. A "w" query, a "w_i" query, and a "w NEAR w_i" query were submitted to the search engine, and the number of matching documents returned was used as an estimate of the probability for the computation of PMI. Therefore, the strength of association of any given word w can be calculated by the following formula:

$$\begin{aligned}\text{SO-PMI}(w) &= \sum_{p_w \in S_p} \text{PMI}(w, p_w) - \sum_{n_w \in S_n} \text{PMI}(w, n_w) \\ &= \sum_{p_w \in S_p} \log_2\left(\frac{hits(w \text{ NEAR } p_w)}{hits(w)hits(p_w)}\right) \\ &\quad - \sum_{n_w \in S_n} \log_2\left(\frac{hits(w \text{ NEAR } n_w)}{hits(w)hits(n_w)}\right)\end{aligned}$$

where $hits()$ denotes the number of matching documents returned by the search engine.

According to the formula, a word w is classified as having a positive semantic orientation if SO-PMI(w) is positive and a negative semantic orientation if SO-PMI(w) is negative. Moreover, the absolute value of SO-PMI(w) can be considered as the strength of the semantic orientation. The authors have tested the SO-PMI method on the HM term set from [17] and the categories Positive and

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

Negative defined in the General Inquirer lexicon [22]. The method achieved an accuracy of 87.13% and 82.84%, respectively, but its performance relied on the size of the corpora indexed by the search engine. Experiments on three corpora of different sizes showed that the accuracy of the method significantly declined with decreasing corpus size [19].

2.2. HAL

HAL was proposed as an implementation of semantic space by Lund and Burgess [28,29]. Semantic space is one in which words are represented by points, and often has a large number of dimensions; the position of each point along each axis is somehow related to the meaning of the word [30]. HAL is enlightened by the intuition phenomenon of the human cognitive process: a human encountering a new concept derives its meaning by employing the accumulated experience of the contexts in which the concept appears.

From Burgess and Lund's research, the procedure of automatically constructing the HAL space from a text corpus can be described as follows. A sliding window of length K is moved across the text corpus at one-word increments, ignoring punctuation and sentence or paragraph boundaries. All words $\underline{w_i, w_{i+1}, w_{i+2}, \cdots w_{i+K-1}}$ within the window are considered as co-occurring with the first word w_i with strengths inversely proportional to the distance between them, i.e., the weight between w_i and w_{i+j} is calculated as $K - j$. After moving the window one word at a time across the whole corpus, the HAL space—an accumulated co-occurrence matrix for all words in the target vocabulary—is produced. The resulting HAL space is an $N \times N$ matrix, where N denotes the vocabulary size. Table 1 presents the HAL space for the example text "*If I get opportunity, I will work hard.*"

As can be seen from Table 1, the HAL space is direction sensitive, i.e., for every word in the target vocabulary, the co-occurrence information for words appearing before/after it is recorded separately by the row/column of the HAL matrix. The row/column pair may be concatenated so that, given a vocabulary T consisting of N words, a word can be represented via a vector of length $2N$ in HAL space. If we neglect the directional sensitivity, and add the row and column into one vector for every row/column pair, the dimension of the representation vector for each word will eventually be reduced to the vocabulary size N. Once the HAL space is built, the context information for each word in vocabulary T will be captured and stored in a HAL vector. As an example, part of the HAL vector for the word "*compete*" is as follows:

$V_{compete}$ = ⟨industry: 105, business: 105, firms: 109, market: 336, effectively: 141, markets: 167, ability: 130, world: 214, better: 117, international: 102, ...⟩.

2.3. Research gaps

Based on our review, there are some research gaps that motivate the work presented in this paper.

Table 1
An example of HAL space ($K = 5$).

	If	I	Get	Opportunity	Will	Work	Hard
If	0	0	0	0	0	0	0
I	7	3	4	5	0	0	0
Get	4	5	0	0	0	0	0
Opportunity	3	4	5	0	0	0	0
Will	1	7	3	4	0	0	0
Work	0	5	2	3	5	0	0
Hard	0	3	1	2	4	5	0

(1) The SO-PMI method is prevented from achieving higher identification accuracy by the very limited number of paradigm words chosen, and by the value of the classification boundary between positive and negative categories being simply set to 0. Thus, SO-PMI is worthy of further study to improve the method by eliminating the impact of the paradigm word selection and by employing a more effective classifier to capture the complex classification rules between the positive category and the negative category.
(2) The time efficiency and applicability of the SO-PMI method is limited because it relies on an Internet search engine. If the Internet is viewed as a massive corpus, a search engine can be viewed as a statistical language model. Therefore, additional research is worthwhile in order to apply a relatively small-scale offline corpus to replace the Internet by employing a more efficient statistical language model, such as HAL, to measure precise co-occurrence information in a corpus.
(3) HAL is a general semantic representation model; however, semantic orientation is a specific type of semantic characteristic. Thus, improving HAL's ability to represent semantic orientation characteristics by adopting a specific base-space is worthy of further study.

3. Method

3.1. S-HAL construction

Following the idea underlying the original HAL, this section presents a specific HAL model that focuses on the representation of semantic orientation information. Our specific HAL model is called Sentiment HAL (S-HAL). During the construction of the original HAL, two questions need to be answered: one is how to set the base-space for characterizing the context, and the other is how to use contextual information to represent the semantic characteristics of words. To answer the first question, the original HAL model adopts the set of all words contained in a corpus as the base-space of the context. To answer the second question, the co-occurrence frequency between a target word and the context is used to express the semantic characteristics of the target word.

To construct S-HAL while focusing on the representation of semantic orientation information, the above two questions need to be considered. As described in Section 2.1, the SO-PMI method shows that the semantic orientation of a word tends to correspond to the semantic orientation of its co-occurring neighbors. Therefore, in constructing S-HAL, we select a specific set of words with definite semantic orientation to use as the base-space of the context, and employ the co-occurrence statistics between target word and context (i.e., the words with definite semantic orientation) to characterize the semantic orientation information of the target word. The construction procedure of S-HAL can be described as follows.

Consider a set of words with definite semantic orientation, denoted by S, where $|S|$ is the size of S. A sliding window of length $2K - 1$ is moved across the text corpus at one-word increments, ignoring punctuation and sentence or paragraph boundaries. All words $\underline{w_{i-K+1}, w_{i-K+2}, \ldots w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \ldots w_{i+K-1}}$ within the window are considered as co-occurring with the target word w_i. For any co-occurrence pair $\langle w_i, w_{i+j}\rangle$, $j \in \{-K+1, -K+2, \ldots, -2, -1, 1, 2, \ldots, K-1\}$, if $w_{i+j} \in S$, the co-occurrence weight between w_i and w_{i+j}, denoted by $n(w_i, w_{i+j})$, is calculated using formula (1) or formula (2). After moving the window over the whole corpus, the S-HAL space, an accumulated co-occurrence weight matrix for all words in the target vocabulary, is produced. The resulting S-HAL space is an $N \times |S|$ matrix, where N denotes the target vocabulary size.

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

There are a variety of computing techniques to evaluate the co-occurrence weight between w_i and w_{i+j} [31]. In this paper, we generate the co-occurrence weight using two strategies: one is a value inversely proportional to distance, similar to the weighting method of the original HAL; the other is a fixed value that ignores the distance factor, similar to the method of counting the total co-occurrence frequency in SO-PMI. The two strategies can be formalized with the following equations.

Strategy 1 (inversely proportional to distance):

$$n(w_i, w_{i+j}) = K - |j| \tag{1}$$

Strategy 2 (ignoring the distance factor):

$$n(w_i, w_{i+j}) = 1 \tag{2}$$

The resulting S-HAL space is an $N \times |S|$ matrix, as shown in (3), wherein each row vector can be considered as the representation of the semantic orientation of word t_i in the target vocabulary.

$$\text{S-HAL} = \begin{bmatrix} s_1 & s_2 & \cdots & s_{|S|} \\ w_{t_1,s_1} & w_{t_1,s_2} & \cdots & w_{t_1,s_{|S|}} \\ w_{t_2,s_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ w_{t_N,s_1} & \cdots & \cdots & w_{t_1,s_1} \end{bmatrix}_{N\times|S|} \tag{3}$$

where $s_i \in S$.

3.2. Semantic orientation identification for terms based on S-HAL

This section presents a method for identifying the WSO of terms based on the classification of semantic orientation feature vectors derived from S-HAL. To build the classification model of semantic orientation feature vectors, standard machine learning techniques are employed. The training set can be derived from the accumulated work of others on semantic orientation research, wherein many hand-labeled sentiment lexicons are available. Specifically, our method consists of the following five steps.

Step 1: For each word in the training set, acquire its semantic orientation feature vector from a pre-trained S-HAL model.
Step 2: Perform feature selection in the original feature space.
Step 3: On the basis of the resulting feature set from Step 2, convert the semantic orientation feature vector into a normalized numeric vector.
Step 4: If phrases are contained in the training set or test dataset, generate the semantic orientation feature vector of phrases based on the semantic orientation feature vector of words contained in the phrase.
Step 5: Train a binary classifier using the normalized numeric vectors of the training set, and then use the resulting classifier to identify the WSO of a given term.

3.2.1. Step 1: Acquire semantic orientation feature vectors

On the basis of the S-HAL model proposed in Section 3.1, this step is straightforward. After setting a base-space S, the S-HAL model can be automatically built on the given corpus. In the S-HAL model, information about each word's semantic orientation is contained in a numeric vector characterized by S, which is exactly the semantic orientation feature vector. Thus, by querying the trained S-HAL model, the semantic orientation feature vector of each word t_i in the training set, denoted by $[w_{t_i a_1}, w_{t_i a_2}, \ldots w_{t_i a_{|S|}}]$ where $a_j \in S$, can be obtained.

3.2.2. Step 2: Perform feature selection for the semantic orientation feature vector

The length of the original semantic orientation feature vector derived from S-HAL is $|S|$, i.e., the size of the base-space. When $|S|$ is a relatively large value, the presence of noisy and redundant features is a major concern. To improve classification accuracy and efficiency, we perform a selection to pick out discriminating features.

In previous studies, Information Gain (IG) has been shown to work well for various text categorization tasks. This is an effective univariate method that considers features individually [32–35]. In this step, we employ IG to conduct a feature selection for the original semantic orientation feature vector. IG can be defined as follows:

$$\begin{aligned} \text{IG}(a_i) = &-\sum_{k=1}^{|C|} P(c_k)\log P(c_k) + P(a_i)\sum_{k=1}^{|C|} P(c_k|a_i)\log P(c_k|a_i) \\ &+ P(\bar{a}_i)\sum_{k=1}^{|C|} P(c_k|\bar{a}_i)\log P(c_k|\bar{a}_i) \end{aligned} \tag{4}$$

where $a_i \in S$, $C = \{positive, negative\}$ denotes the set of categories, $P(c_k)$ denotes the probability that category c_k occurs, $P(a_i)$ denotes the probability that feature a_i occurs, and $P(\bar{a}_i)$ denotes the probability that feature a_i does not occur.

3.2.3. Step 3: Convert the semantic orientation feature vector into a normalized numeric vector

Following feature selection, the semantic orientation feature vector is still frequency-based, i.e., a highly frequent feature generally receives a high weight. To achieve better classification accuracy, the semantic orientation feature vectors need to be re-weighted and normalized by appropriate methods. In this step, we re-weight and normalize the semantic orientation feature vectors by one of the following methods.

3.2.3.1. Re-weighting method I: tf.idf scheme. This method adopts a weighting scheme analogous to TF-IDF [36] and the cosine formula to re-weight and normalize the dimensions of the semantic orientation feature vector, as shown below:

$$w_{t_i a_j} = w_{t_i a_j} * \log\frac{N_{vector}}{vf(a_j)} \tag{5}$$

$$w_{t_i a_j} = \frac{w_{t_i a_j}}{\left(\sum_{l=1}^{\dot{N}} (w_{t_i a_l})^2\right)^{\frac{1}{2}}} \tag{6}$$

where N_{vector} denotes the total number of vectors, $vf(a_j)$ denotes the number of vectors with the dimension a_j, and $\dot{N}$ denotes the size of the feature subset generated by feature selection.

3.2.3.2. Re-weighting method II: PMI scheme. Considering that the SO-PMI method is a very effective approach to measuring semantic orientation, a weighting scheme analogous to PMI calculation [18,19,25,37] is employed to re-weight the dimensions of the semantic orientation feature vector. This method can be described as:

$$w_{t_i a_j} = \frac{w_{t_i a_j}}{\sum_{l=1}^{\dot{N}} w_{t_i a_l} \cdot \sum_{k=1}^{\dot{N}} w_{a_j a_k}} \tag{7}$$

3.2.3.3. Re-weighting method III: 0–1 scheme. Considering the time efficiency of classifier training, a simple re-weighting method, the 0–1 scheme, is designed as follows:

$$w_{t_i a_j} = \begin{cases} 1 & \text{if } w_{t_i a_j} > 0 \\ 0 & \text{if } w_{t_i a_j} = 0 \end{cases} \tag{8}$$

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

After each dimension is re-weighted by one of above methods, all semantic orientation feature vectors are converted into normalized numeric vectors that can be used for various machine learning algorithms.

3.2.4. Step 4: Combine the semantic orientation feature vectors of phrases

From the definition of S-HAL, the semantic orientation feature vector of phrases cannot be directly queried from the model. When phrases are contained in the training or test sets, the semantic orientation feature vector of phrases must be generated by heuristically combining those of words in the phrase. In this step, we acquire the semantic orientation feature vector of phrases by using one of the following heuristic combination methods.

Let $t_1 \oplus t_2$ denote the phrase consisting of words t_1 and t_2, and $w_{t_1 \oplus t_2 a_j}$ denote the weight of feature a_j of the semantic orientation feature vector of phrase $t_1 \oplus t_2$. We have:

Combination method I: Min scheme

$$w_{t_1 \oplus t_2 a_j} = \mathrm{Min}(w_{t_1 a_j}, w_{t_2 a_j}) \tag{9}$$

Combination method II: Max scheme

$$w_{t_1 \oplus t_2 a_j} = \mathrm{Max}(w_{t_1 a_j}, w_{t_2 a_j}) \tag{10}$$

Combination method III: Product scheme

$$w_{t_1 \oplus t_2 a_j} = w_{t_1 a_j} \cdot w_{t_2 a_j} \tag{11}$$

Obviously, the resulting phrase is a new word, which, in turn, can be composed into other phrases via heuristic combination.

3.2.5. Step 5: Train a binary classifier for semantic orientation identification

After Steps 1–4 (if datasets do not contain phrases, this step ought to be skipped) are finished, the semantic orientation feature vector of each term in the training set is converted into a normalized numeric vector. In this step, these vectors are input into a standard supervised learner that generates a binary classifier of semantic orientation.

SVMs have been shown to be highly effective at traditional text categorization. In prior studies on semantic orientation classification, the SO-A method can be viewed as a simple linear binary classifier that is, in a sense, similar to a SVM classifier. Indeed, a SVM classifier exhibited the best performance in Esuli and Sebastiani's work [12]. Therefore, we employ a SVM classifier to train the model to identify the semantic orientation of terms.

Note that SVMs are based on the structural risk minimization principle from computational learning theory, i.e., the basic idea behind the SVM training procedure is to seek a hyperplane that separates the elements in one class from those in another class with as large a margin as possible. According to the analysis introduced in Section 2.3, one of the shortcomings of the SO-PMI method is that the classification boundary is manually chosen as the value 0, and one of the aims of this work is to optimize the classification boundary. Therefore, in principle, a SVM classifier is helpful for achieving our goal of an optimal classification boundary.

In all experiments reported in this paper, the libsvm package is used for training and testing the SVM classifier, with the RBF kernel adopted and all other parameters set to their default values (the libsvm package is available at http://www.csie.ntu.edu.tw/~cjlin/).

4. Experiments

In order to evaluate the proposed method, we conducted three different experiments. The first experiment was a comparison of the proposed method with other known methods on the same test bed. In the second experiment, we analyzed the effect of the model parameters' configuration on the identification performance. The final experiment mainly evaluated the heuristic combination method for producing the semantic orientation feature vector of phrases, which is crucial for identifying the semantic orientation of phrases. This section details the experimental datasets, implementation, and results.

4.1. Data sets and evaluation methodology

4.1.1. Corpora and lexicons

The experimental evaluation was performed on two different Chinese corpora (used for training the S-HAL model) and on three different Chinese sentiment lexicons (used for training and testing the classifier). The Chinese text POS tool of ICTCLAS was applied to parse and tag the parts-of-speech in all corpora (the ICTCLAS package is available at http://ictclas.org/ictclas_download.asp).

To construct the S-HAL model, two versions of the *Sogou CS corpus* were employed (complete version *SogouCS corpus* and reduced version *SogouCSReduced corpus*; the Sogou CS corpus is released by Sogou laboratory and is available at http://www.sogou.com/labs/dl/cs.html). The *SogouCS corpus* is the set of all news pages on www.sohu.com from January 2008 to June 2008. The *SogouCS corpus* contained a total of 2,820,059 pages and approximately 530 million words after removing all stop words and numerical symbols. After further removing all infrequent words that occurred fewer than 40 times in the corpus, the *SogouCS corpus* contained a total of 116,233 distinct words, which formed the target vocabulary of S-HAL, denoted by *V*. As a subset of *SogouCS corpus*, the *SogouCSReduced corpus* is the set of news pages on www.sohu.com in June 2008. The *SogouCSReduced corpus* contained a total of 576,364 pages and approximately 120 million words after removing all stop words and numerical symbols.

To train and test the SVM classifier, we adopted three different lexicons that were hand-labeled with a semantic orientation (positive or negative): *H lexicon*, *T lexicon*, and *P lexicon*. The *H lexicon* was a list of words with definite semantic orientation, released by Hownet [38] (the *H* lexicon is available at www.keenage.com/html/c_index.html); the *T lexicon* was a list of words with definite semantic orientation created by Li [39] (the *T* lexicon is available at http://nlp.csai.tsinghua.edu.cn/site/index.php?page=resources); the *P lexicon* was a list of phrases that were automatically extracted from user comments in Web forums and manually labeled with a semantic orientation by ourselves.

The original *H lexicon* consisted of 4528 words with positive semantic orientation and 4320 words with negative semantic orientation. After removing all words for which the semantic orientation labels were self-contradictory or contradictory to the *T lexicon*, there were 4366 words with positive semantic orientation and 4178 words with negative semantic orientation. After further removing all infrequent words that were not contained in the target vocabulary of S-HAL, 3172 words with positive semantic orientation and 2534 words with negative semantic orientation were left. Ultimately, this list of 5706 words was used as the *H lexicon* in our experiments.

The original *T lexicon* consisted of 5567 words with positive semantic orientation and 4468 words with negative semantic orientation. After preprocessing in a similar way to the *H lexicon*, a list of 6850 words (3986 positive and 2864 negative) was ultimately used as the *T lexicon* in our experiments.

Note that some 2074 words overlapped between the *H lexicon* and the *T lexicon*. Thus, there were a total of 10,482 distinct words when these were merged into a larger lexicon. In the experiments, we denote this larger list as the *H + T lexicon*, and use it as the initial base-space of S-HAL.

The *P lexicon* consisted of 920 phrases with positive semantic orientation and 864 phrases with negative semantic orientation.

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

All phrases were extracted from consumer reviews or news comments in Web forums, and were manually labeled with majority rule by five computer science postgraduates. Moreover, all words comprising these phrases were contained in the target vocabulary of S-HAL.

4.1.2. Dataset construction

On the basis of the three lexicons described above, we constructed the six datasets below for training and testing the SVM classifier.

(1) Data set 1 was based on the *H + T lexicon*. We performed fivefold cross-validation on all 10,482 words in the *H + T lexicon*, and balanced these words across classes.
(2) Data set 2 was based on the *H lexicon*. We performed fivefold cross-validation on all 5706 words in the *H lexicon*, and balanced these words across classes.
(3) Data set 3 was based on the *T lexicon*. We performed fivefold cross-validation on all 6850 words in the *T lexicon*, and balanced these words across classes.
(4) Data set 4 was based on the *H lexicon* and the *T lexicon*. After 2074 words that overlapped both the *H lexicon* and the *T lexicon* were removed from the *H lexicon*, the remaining 3632 words in the *H lexicon* were used as a training set. The *T lexicon* was also used as a testing set.
(5) Data set 5 was based on the *H lexicon* and the *T lexicon*. After 2074 words that overlapped both the *H lexicon* and the *T lexicon* were removed from the *T lexicon*, the remaining 4776 words in the *T lexicon* were used as a training set. The *H lexicon* was also used as a testing set.
(6) Data set 6 was based on the *H + T lexicon* and the *P lexicon*. In this data set, the *H + T lexicon* was used as a training set, and the *P lexicon* was used as a testing set.

4.1.3. Evaluation measures

To evaluate the performance of our method, the standard *Accuracy* and *F-measure* were used. The *Accuracy* is defined simply as the ratio of number of correctly classified words to the total number of words.

The *F-measure* is computed by combining the *Precision* and *Recall* in the following way:

$$F\text{-}measure = 2 \cdot Precision \cdot Recall/(Precision + Recall)$$

where *Precision* is defined as the ratio of number of correctly assigned category *C* words to the total number of words classified as category *C*. *Recall* is the ratio of correctly assigned category *C* words to the total number of words actually in category *C*. Since *F-measure* is computed for each category separately, for a comprehensive evaluation, we aggregated the *F-measure* scores over two categories by using the Macro- and Micro-averages of the *F-measure* scores. In Micro-average, each word is given an equal weight, while Macro-average gives each category equal weight.

In addition, to verify the impact of different parameter settings on the classifier's performance, we employed McNemar's significance test [40]. This is a χ^2-based significance test under the null hypothesis that two classifiers will have the same error rate on the test set. The statistic χ is defined as:

$$\chi = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10})$$

where n_{01} denotes the number of examples misclassified by classifier A but not by classifier B and n_{10} denotes the number of examples misclassified by classifier B but not by classifier A. Dietterich showed that, under the null hypothesis, χ is approximately distributed as a χ^2 distribution with one degree of freedom. In this paper, we selected the significance level 0.05 to correspond to the threshold $\chi = 3.84$.

4.2. Experiment 1: Comparison of proposed method against other known methods

In order to evaluate the effectiveness of the proposed method for semantic orientation identification, we conducted experiments on Data sets 1–6. For comparison, the SO-PMI method [19] and the methods presented in [12,27] were also implemented to classify the same data sets. The SO-PMI method was regarded as one of the best approaches in previous research on identifying the semantic orientation of terms; the method presented in [12] not only showed excellent performance but was also used as the technical basis for generating SentiWordNet; the method presented in [27] is the latest published method for identifying the semantic orientation of Chinese words.

In implementing the proposed method, we adopted the baseline parameter configuration. Specifically, the *SogouCS* corpus was adopted for training S-HAL; the target vocabulary of S-HAL was the set *V* defined in Section 4.1; the list of 10,482 words contained in the *H + T lexicon* was used as the initial base-space of S-HAL; the sliding window for training S-HAL had a length of 10 words; the strategy for evaluating co-occurrence weights was Strategy 1 defined in Section 3.1; the feature selection step for the semantic orientation vector was skipped; the re-weighting method for normalizing the semantic orientation vector was method I (*tf.idf* scheme) defined in Section 3.2; and for phrases in Data set 6, combination method I (Min scheme) defined in Section 3.2 was used to generate the semantic orientation vector of a phrase.

To implement the SO-PMI method, we adopted the *Yahoo! Search API* to obtain *hit*() values (the Yahoo! Search API is available at http://developer.yahoo.com/search/). In our experiment, we used the NEAR operator provided by the Yahoo! API to estimate the co-occurrence frequency, and the scope for the search engine was set to "in the web." In selecting paradigm words for the SO-PMI method, we employed the two groups of words listed in Table 2. One consisted of 14 words from the original SO-PMI method [19] and the other was composed of 40 words based on the work in [41].

The method presented in [12] was implemented with the paradigm words in Table 2 as the seed sets for expansion, and synonym and antonym expansion was simultaneously performed for five iterations based on the lexical relation source provided by an electronic thesaurus (the thesaurus is available at www.365zn.com/fyc/ and www.ir-lab.org/). The creation of textual representations of terms is based on the use of glosses extracted from *The Contemporary Chinese Dictionary*. In the process of training a text classifier, a SVM binary classifier was adopted.

To perform the method presented in [27], the *SogouCS* corpus was adopted to compute the co-occurrence similarity, and other experimental settings were consistent with the original method in [27].

Table 3 shows the results for our method and other methods across Data sets 1–6. Comparison of the results shows that our method outperforms the other methods across all data sets, although the results achieved with different data sets varies greatly. Conceptually, our approach is most similar to the SO-PMI method because the co-occurrence information in massive corpora is used as the key feature for identifying the semantic orientation of terms. Experimental results suggest that our approach makes more efficient use of these corpora than the SO-PMI method. Similar to the method presented in [12], our approach trains a binary classifier to identify semantic orientation based on the feature vector representation of terms. However, the difference is that our approach adopts co-occurrence information rather than semantic

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

Table 2
Paradigm words used in the SO-PMI method and the method presented in [12].

Group	Positive		Negative	
14 word version	Good (好)	Nice (美好)	Bad (坏)	Nasty (讨厌)
	Excellent (杰出)	Positive (积极)	Poor (卑鄙)	Negative (消极)
	Fortunate (幸运)	Correct (正确)	Unfortunate (不幸)	Wrong (错误)
	Superior (优秀)		Inferior (低劣)	
40 word version	Grateful (感激)	Splendid (出色)	Distorted (扭曲)	Wrong (错误)
	Energetic (活力)	Brilliant (光辉)	Corrupted (崩溃)	Agonizing (烦恼)
	Mellow (成熟)	Virtuous (善良)	Psychopathic (变态)	Poor (卑鄙)
	Remarkable (非凡)	Beautiful (漂亮)	Morose (郁闷)	Negative (消极)
	Excellent (杰出)	Honest (诚实)	Dishonest (虚假)	Nasty (讨厌)
	Positive (积极)	Fortunate (幸运)	Hidebound (保守)	Tragic (悲惨)
	Polite (礼貌)	Harmonious (和谐)	Rude (粗暴)	Horrible (可怕)
	Nice (美好)	Comfortable (舒服)	Unfortunate (不幸)	Stupid (痴呆)
	Lenient (宽容)	Peaceful (和平)	Inferior (低劣)	Unhealthy (不良)
	Correct (正确)	Superior (优秀)	Silly (糊涂)	Shameful (耻辱)

Table 3
Accuracy, Macro-F1, and Micro-F1 results of the proposed method using the baseline configuration, and comparison with the performance of the SO-PMI method.

Method		Data set 1	Data set 2	Data set 3	Data set 4	Data set 5	Data set 6
Proposed method	Accuracy	0.8646	0.8503	0.9241	0.8711	0.8106	0.9211
	Macro-F1	0.8629	0.8484	0.9221	0.8691	0.8074	0.9189
	Micro-F1	0.8648	0.8503	0.9242	0.8718	0.8102	0.9206
SO-PMI (with 14 paradigm words)	Accuracy	0.7453	0.7143	0.7790	0.7790	0.7143	0.6809
	Macro-F1	0.7429	0.7092	0.7788	0.7788	0.7092	0.6475
	Micro-F1	0.7399	0.7049	0.7777	0.7777	0.7049	0.6477
SO-PMI (with 40 paradigm words)	Accuracy	0.8304	0.8006	0.8696	0.8696	0.8006	0.7506
	Macro-F1	0.8275	0.7976	0.8661	0.8661	0.7976	0.7435
	Micro-F1	0.8303	0.8003	0.8696	0.8696	0.8003	0.7436
Method in [12] (with 14 paradigm words)	Accuracy	0.8178	0.7861	0.8593	0.8593	0.7861	–
	Macro-F1	0.8122	0.7834	0.8537	0.8537	0.7834	–
	Micro-F1	0.8169	0.7857	0.8585	0.8585	0.7857	–
Method in [12] (with 40 paradigm words)	Accuracy	0.8142	0.7880	0.8583	0.8583	0.7880	–
	Macro-F1	0.8102	0.7812	0.8529	0.8529	0.7812	–
	Micro-F1	0.8146	0.7867	0.8585	0.8585	0.7867	–
Method in [27]	Accuracy	0.7589	0.7057	0.7725	0.7725	0.7057	–
	Macro-F1	0.7572	0.7012	0.7689	0.7689	0.7012	–
	Micro-F1	0.7586	0.7044	0.7716	0.7716	0.7044	–

"–" Denotes that the corresponding method cannot be used for phrase-level classification.

information of a glossary. Experimental results suggest that co-occurrence information in a massive corpus is more efficient than semantic information from a glossary in the semantic orientation identification of terms. Compared with their performance on Data sets 1 and 4, our method and all comparison methods showed a relatively low performance on Data sets 2 and 3. This may be because the *H lexicon* contains more noise than does the *T lexicon*. We also noticed that the results achieved by the SO-PMI method with 14 paradigm words were clearly inferior to the results reported for the English task in [19]. After the set of paradigm words was expanded from 14 to 40 words based on related research in the Chinese field, the performance of the SO-PMI method greatly improved. Its time efficiency decreased linearly with the increase in the size of the paradigm words set. In our experiment, the time cost of classifying each word was 260 s using the SO-PMI method with 40 paradigm words. As a comparison, in our method, the most time-consuming work is to build the S-HAL model, but it is one-time cost. Once S-HAL was constructed, the time cost for training the classifier was about 900 s, and the time cost of classifying each word was 0.12 s based on the trained classifier.

4.3. Experiment 2: Effect of parameter configuration

In this experiment, we analyzed the effect of the model parameters' configuration on the identification performance.

4.3.1. Varying the number of dimensions of the semantic orientation feature vector

As mentioned in Section 3.2, we performed IG-based feature selection to pick out discriminating features. In this section, we explore the effect of varying the size of the feature set.

In the experiment, we evaluated 13 different feature set sizes by manually setting IG thresholds. To convert the semantic orientation feature vector into a normalized numeric vector, the three different re-weighting methods introduced in Section 3.2 were used. The other setup used in the experiment was the baseline configuration introduced in Section 4.2.

Fig. 1 displays the accuracy of the three re-weighting methods with respect to feature number. This shows that the *tf.idf* and 0–1 re-weighting schemes outperform the PMI re-weighting scheme in almost all cases. However, the effect of the *tf.idf* re-weighting scheme and the 0–1 re-weighting scheme was very similar. Thus, we employed McNemar's significance test to further evaluate the effect of these two schemes, and the results are presented in Table 4. In addition, we can observe from Fig. 1 that once the number of features exceeded 6000, the performance of each method improved only slightly with a further increase in the number of features.

As Table 4 shows, at the test level of $p = 0.05$, the two re-weighting schemes (*tf.idf* and 0–1) have the same performance across all data sets when the size of the feature set employed exceeds 1000.

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

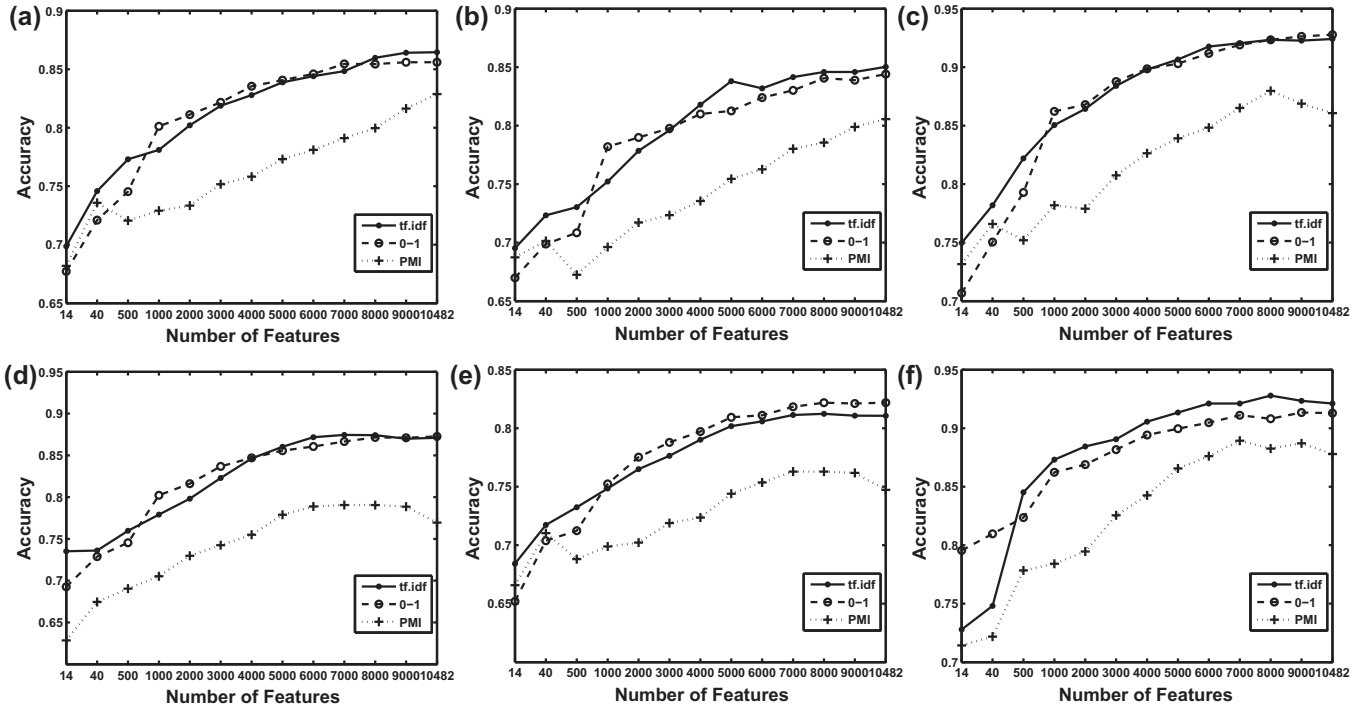

Fig. 1. Effect of varying the number of dimensions of the semantic orientation feature vector for: (a) Data set 1; (b) Data set 2; (c) Data set 3; (d) Data set 4; (e) Data set 5; (f) Data set 6.

Table 4
P-values of McNemar's significance test for difference between the *tf.idf* and 0–1 re-weighting schemes.

	Size of feature set												
	14	40	500	1000	2000	3000	4000	5000	6000	7000	8000	9000	10482
Data set 1	<0.05	<0.05	<0.05	<0.05	–	–	–	–	–	–	–	–	–
Data set 2	–	–	–	<0.05	–	–	–	–	–	–	–	–	–
Data set 3	<0.05	<0.05	<0.05	–	–	–	–	–	–	–	–	–	–
Data set 4	<0.05	–	–	<0.05	–	–	–	–	–	–	–	–	–
Data set 5	<0.05	–	<0.05	–	–	–	–	–	–	–	–	–	–
Data set 6	<0.05	<0.05	<0.05	–	–	–	–	–	–	–	–	–	–

"–" Denotes *p*-value > 0.05.

Although there is a significant performance difference when employed in a relatively small feature set, such sets are rarely used in practice due to their poor performance. Thus, we believe that the *tf.idf* re-weighting scheme is not superior to the 0–1 re-weighting scheme.

In addition, Macro-F1 and Micro-F1 also exhibited similar characteristics to Accuracy. Hence, for reasons of space, the rest of this paper omits to report these measures in detail.

4.3.2. Varying the training set size for the SVM classifier

For supervised learning algorithms, such as SVM, the size of the training set may affect the performance of the classifier. In the previous experiments, all SVM classifiers were trained by sets of fixed size. In this section, we will explore the effect of varying the size of the training set.

This experiment was conducted on Data sets 1–3, each of which was randomly split into a training set and a testing set according to different specified proportions. The proportions adopted in the experiment were 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, and 1:9. In order to eliminate the impact of random factors on the results, we repeated the experiment 10 times for each split proportion and computed the average results. We adopted the three different re-weighting methods introduced in Section 3.2 to convert the semantic orientation feature vector into a normalized numeric vector, and the other setup used in this experiment was the baseline configuration introduced in Section 4.2.

Fig. 2 shows the accuracy curves with respect to varying sizes of training set for the SVM classifier. This figure shows that, by adopting either the *tf.idf* re-weighting scheme or the 0–1 re-weighting scheme, the number of training samples has little effect on the performance of the SVM classifier. However, if the PMI re-weighting scheme is used, the performance of the SVM classifier would be greatly affected by the number of training samples.

4.3.3. Varying the parameter setup for training S-HAL

As discussed in Section 3.2, deriving a vector representation of the semantic orientation information of words from S-HAL is a crucial step. The experiment in this section is designed to explore the effect of the S-HAL model on the result of semantic orientation identification. We tested various parameters, including corpus size, length of sliding window, and weighting strategy, in training S-HAL.

This experiment was conducted on Data set 1. In the process of constructing S-HAL, we experimented with two corpora (*SogouCS* and *SogouCSReduced*), sliding window lengths varying from 4 to 14 words, and the two weighting strategies introduced in Section

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

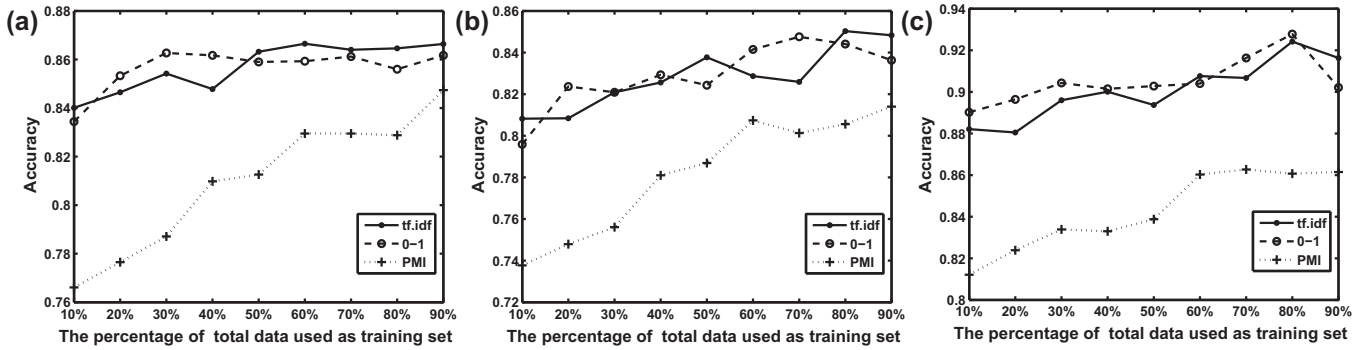

Fig. 2. Effect of varying the training set size for the SVM classifier on: (a) Data set 1; (b) Data set 2; (c) Data set 3.

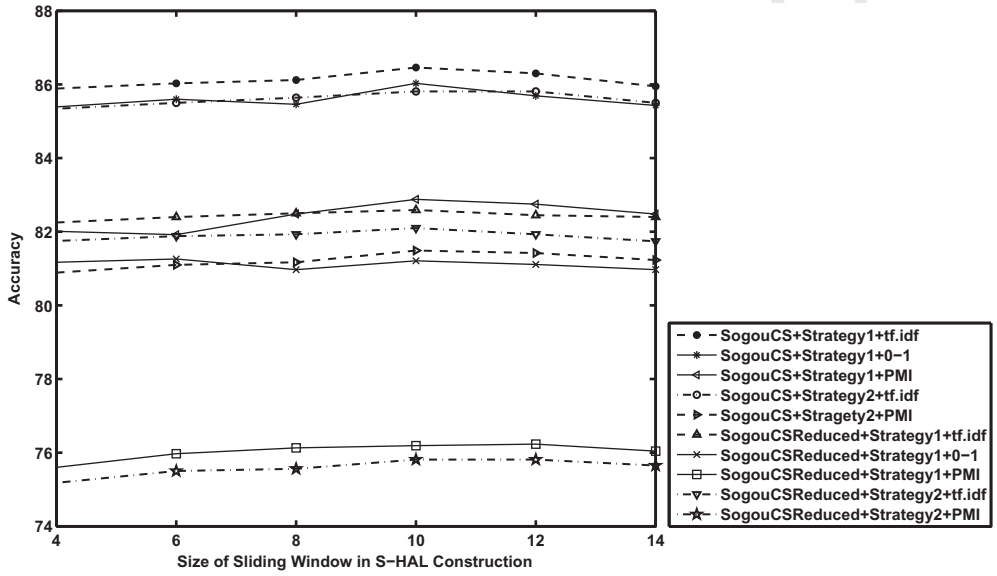

Fig. 3. Effect of varying the parameter setup for training S-HAL.

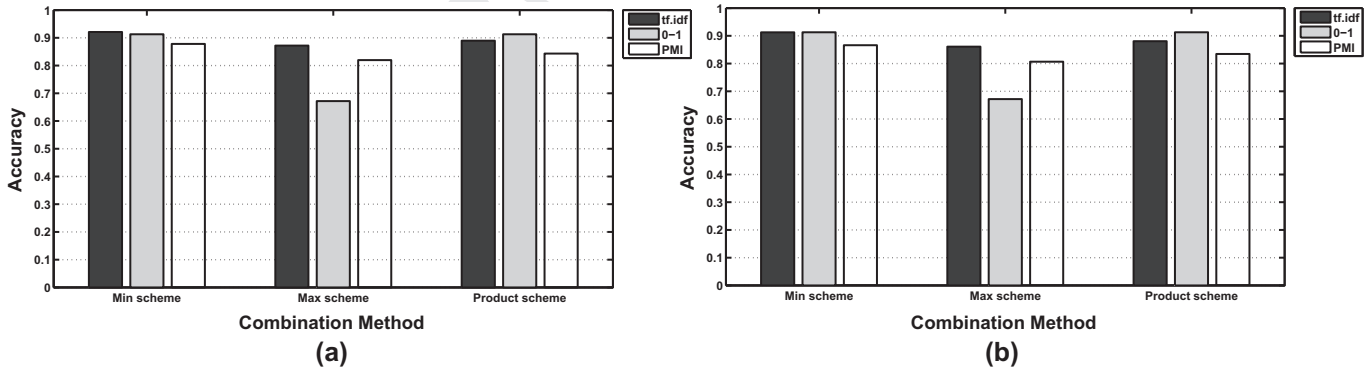

Fig. 4. Effect of varying the combination method. (a) Co-occurrence weighting strategy 1 adopted for training S-HAL. (b) Co-occurrence weighting strategy 2 adopted for training S-HAL.

3.1. The three re-weighting methods introduced in Section 3.2 were adopted to convert the semantic orientation feature vector into a normalized numeric vector, and the other setup used in this experiment was the baseline configuration introduced in Section 4.2.

Fig. 3 illustrates the accuracy of the semantic orientation identification algorithm for different S-HAL training parameter configurations. Fig. 3 shows that corpus size is a crucial factor in improving the performance of the present method. Another observation is that, although a 10-word sliding window is an appropri-

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011

ate choice, the length of this window does not have a significant effect on the accuracy of the identification algorithm. In addition, weighting strategy 1 slightly outperforms weighting strategy 2 for all re-weighting schemes.

4.4. Experiment 3: Effect of the combination method

As discussed in Section 3.2, it is necessary to generate semantic orientation feature vectors by heuristic combination when phrases are contained in the training or test data. In Section 3.2, three combination methods were proposed. In this section, we will investigate which combination method is the most effective for identifying the semantic orientation of phrases.

In this experiment, Data set 6 was employed, and the three re-weighting methods introduced in Section 3.2 were adopted to convert semantic orientation feature vectors into normalized numeric vectors. To train the S-HAL model, we experimented with the *SogouCS* corpus and the two weighting strategies introduced in Section 3.1. The other setup used in this experiment was the baseline configuration introduced in Section 4.2.

Fig. 4 plots the performance of the three combination methods for semantic orientation identification. This shows that the Min scheme has a clear advantage in phrase combination accuracy for each of the weighting strategies of S-HAL. Note that the performance of the Max scheme dramatically decreased when the 0–1 re-weighting method was adopted. This may be because the 0–1 re-weighting method enlarges the effect of information redundancy brought about by the Max combination scheme. Besides, compared with co-occurrence weighting strategy 1 used for training S-HAL, co-occurrence weighting strategy 2 shows relatively low performance across all combination methods and re-weighting methods. This is consistent with the experimental results in Section 4.3.3.

5. Conclusion

In this study, we have presented a novel method for automatically identifying the semantic orientation of subjective terms (words or phrases). We first proposed a semantic orientation representation model that integrated the ideas underlying the HAL model and the SO-PMI method, and then, on the basis of the model and freely available sentiment lexicons, a binary classifier was trained to predict the semantic orientation of any given term. We conducted experiments to compare our method with other known methods, and investigated the effect of varying the configuration of some principal model parameters.

Experimental results indicated that our method outperformed the SO-PMI method and several other published methods. Thus, it will facilitate a wide variety of applications in sentiment analysis due to its high accuracy and ability to be used without the online support of the Internet. Moreover, compared with the original HAL model, S-HAL showed advantages in modeling semantic orientation characteristics. As an enhanced model that transforms semantic space into a subspace focusing on semantic orientation information, S-HAL not only provides a more accurate representation of semantic orientation characteristics, but also confirms that a more specific semantic-subspace model can be developed by employing an appropriate base-space and corresponding measurement. Furthermore, our work examined the hypothesis proposed in SO-PMI that the semantic orientation of a word tends to correspond to that of its co-occurring neighbors, and revealed that different methods for weighting co-occurrence strength have a significant effect on semantic orientation identification.

Finally, it is important to point out that, like many existing methods [12,17,19,26,27], our approach assumes that the classification terms are subjective. Determining whether a term is subjective or objective is a separate issue. An important direction for future work is to apply the S-HAL model to the Subjective/Objective classification of terms. In addition, as all experiments in this paper were performed on Chinese corpora and lexicons, it would also be interesting to test the effect of the present method on English corpora and lexicons in future work.

Acknowledgments

This work was jointly supported by the National Natural Science Foundation of China (Grant Numbers 61173111 and 60774086) and the Ph.D. Programs Foundation of Ministry of Education of China (Grant Number 20090201110027).

References

[1] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, 2002, pp. 79–86.
[2] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in: Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003, pp. 519–528.
[3] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 2004, pp. 168–177.
[4] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005, pp. 347–354.
[5] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (1–2) (2008) 1–135.
[6] E. Boiy, M.F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Information Retrieval 12 (5) (2009) 526–558.
[7] B. Liu, Sentiment analysis and subjectivity, in: N. Indurkhya, F.J. Damerau (Eds.), Handbook of Natural Language Processing, 2010.
[8] H. Chen, D. Zimbra, AI and opinion mining, IEEE Intelligent Systems 25 (3) (2010) 74–80.
[9] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment in short strength detection informal text, Journal of the American Society for Information Science and Technology 61 (12) (2010) 2544–2558.
[10] A. Neviarouskaya, H. Prendinger, M. Ishizuka, SentiFul: a lexicon for sentiment analysis, IEEE Transactions on Affective Computing 2 (1) (2011) 1–15.
[11] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, Computational Linguistics 37 (1) (2011) 9–27.
[12] A. Esuli, F. Sebastiani, Determining the semantic orientation of terms through gloss classification, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, 2005, pp. 617–624.
[13] M.P. O'Mahony, B. Smyth, A classification-based review recommender, Knowledge-Based Systems 23 (4) (2010) 323–329.
[14] C. Kaiser, S. Schlick, F. Bodendorf, Warning system for online market research – identifying critical situations in online opinion formation, Knowledge-Based Systems 24 (2011) 824–836.
[15] S. Cleger-Tamayo, J.M. Fernandez-Luna, J.F. Huete, Top-N news recommendations in digital newspapers, Knowledge-Based Systems 27 (2012) 180–189.
[16] R.T. Sikora, K. Chauhan, Estimating sequential bias in online reviews: a Kalman filtering approach, Knowledge-Based Systems 27 (2012) 314–321.
[17] V. Hatzivassiloglou, K.R. McKeown, Predicting the semantic orientation of adjectives, in: Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, Madrid, Spain, 1997, pp. 174–181.
[18] P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 79–86.
[19] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 21 (4) (2003) 315–346.
[20] H. Takamura, T. Inui, M. Okumura, Extracting semantic orientations of words using spin model, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 133–140.
[21] H.D. Lasswell, J.Z. Namenwirth, The Lasswell Value Dictionary, Yale University Press, New Haven, 1969.
[22] P.J. Stone, D.C. Dunphy, M.S. Smith, D.M. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis, MIT Press, Cambridge, MA, 1966.
[23] G.A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.

[24] A. Esuli, F. Sebastiani, SENTIWORDNET: a publicly available lexical resource for opinion mining, in: Proceedings of the 5th Conference on Language Resources and Evaluation, 2006, pp. 417–422.
[25] S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the 7th Conference on Language Resources and Evaluation, 2008, pp. 2200–2204.
[26] J. Kamps, M. Marx, R.J. Mokken, M.D. Rijke, Using WordNet to measure semantic orientations of adjectives, in: Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004, vol. 4, pp. 1115–1118.
[27] W. Du, S. Tan, Optimizing modularity to identify semantic orientation of Chinese words, Expert Systems with Applications 37 (7) (2010) 5094–5100.
[28] K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence, Behavior Research Methods 28 (2) (1996) 203–208.
[29] C. Burgess, K. Livesay, K. Lund, Explorations in context space: words, sentences, discourse, Discourse Processes 25 (2) (1998) 211–257.
[30] C.E. Osgood, G.J. Suci, P.H. Tannenbaum, The Measurement of Meaning, University Illinois Press, Oxford, England, 1957.
[31] M. Sahlgren, The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces, Ph.D. Thesis, Stockholm University, 2006.
[32] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, Springer, 1998, pp. 137–142.
[33] G. Forman, An extensive empirical study of feature selection metrics for text classification, The Journal of Machine Learning Research 3 (2003) 1289–1305.
[34] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Transactions on Information Systems 26 (3) (2008) 1–34.
[35] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, IEEE Transactions on Knowledge and Data Engineering 23 (3) (2010) 447–462.
[36] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management 24 (5) (1988) 513–523.
[37] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational Linguistics 16 (1) (1990) 22–29.
[38] Z. Dong, Q. Dong, HowNet and the Computation of Meaning, World Scientific Publishing Co., Inc., River Edge, NJ, 2006.
[39] J. Li, M. Sun, Experimental study on sentiment classification of Chinese review using machine learning techniques, in: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, 2007, pp. 393–400.
[40] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895–1923.
[41] Y. Zhu, J. Min, Y. Zhou, X. Huang, L. Wu, Semantic orientation computing based on HowNet, Journal of Chinese Information Processing 20 (1) (2006) 14–20.

Please cite this article in press as: T. Xu et al., Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.04.011